LEGaTO: Towards Energy-Efficient, Secure, Fault-tolerant Toolset for Heterogeneous Computing

Adrian Cristal, Osman S. Unsal, Xavier Martorell, Paul Carpenter, Raul De La Cruz, Leonardo Bautista, Daniel Jimenez, Carlos Alvarez, Behzad Salami, Sergi Madonar (BSC), Miquel Pericàs, Pedro Trancoso (Chalmers), Micha vor dem Berge, Gunnar Billung-Meyer, Stefan Krupop, Wolfgang Christmann (CHR), Frank Klawonn, Amani Mihklafi (HZI), Tobias Becker, Georgi Gaydadjiev (Maxeler), Hans Salomonsson, Devdatt Dubhashi (MIS), Oron Port, Yoav Etsion (Technion), Vesna Nowack, Christof Fetzer (TUD), Jens Hagemeyer, Thorsten Jungeblut, Nils Kucza, Martin Kaiser, Mario Porrmann (UBI), Marcelo Pasin, Valerio Schiavoni, Isabelly Rocha,Christian Göttel, Pascal Felber (UniNE)

ABSTRACT

LEGaTO is a three-year EU H2020 project which started in December 2017. The LEGaTO project will leverage task-based programming models to provide a software ecosystem for Made-in-Europe heterogeneous hardware composed of CPUs, GPUs, FPGAs and dataflow engines. The aim is to attain one order of magnitude energy savings from the edge to the converged cloud/HPC.

1 INTRODUCTION

In the last couple of decades, technological advances in the ICT sector have been the dominant factors in global economic growth, not to mention an increase in the quality of life for billions of people. At the heart of this advance lies Moore's Law, which states that the number of transistors in an integrated chip will double every 18 to 24 months with each step in the silicon manufacturing technology node. However, due to fundamental limitations of scaling at the atomic scale, coupled with heat density problems of packing an ever-increasing number of transistors in a unit area, Moore's Law has slowed down in the last two years or so and will soon stop altogether [1]. The implication is that, in the future, the number of transistors that could be incorporated in a processor chip will not increase. This development threatens the future of the ICT sector as a whole. As a solution to this challenge, there recently have been dramatically increased efforts toward heterogeneous computing, including integration of heterogeneous cores on die (ARM), utilizing general-purpose GPUs (NVIDIA), combining CPUs and GPUs on the same die (Intel, AMD, ARM), leveraging FPGAs (Altera, Xilinx), integrating CPUs with FPGAs (Xilinx), and coupling FPGAs and CPUs in the same package (IBM–Altera, Intel–Altera). Heterogeneity aims to solve the problems associated with the end of Moore's Law by incorporating more specialized compute units in the system hardware and by utilizing the most efficient compute unit for each computation. However, while software-stack support for heterogeneity is relatively well developed for performance, it is severely lacking for power- and energy-efficient computing. Given that the ICT sector is responsible for ~5% of global electricity consumption [2], software stack support for energy-efficient heterogeneous computing is critical to the future growth of the ICT industry. The primary ambition of the LEGaTO project is to address this challenge by starting with a Made-in-Europe mature software stack and by optimizing this stack to support energy-efficient computing on a commercial cutting-edge European-developed CPU–GPU–FPGA heterogeneous hardware substrate [3] and FPGA-based Dataflow Engines (DFE), which will lead to an order of magnitude increase in energy efficiency. The LEGaTO project will utilize a completely integrated software system stack supporting generalized tasks for low-energy, secure and reliable parallel computing. We foresee that optimization opportunities for low-energy computing can be maximized through the task abstraction.

The mature software stack that will be the baseline for development of the project is a task-based programming model family with a dataflow-based runtime. These task-based programming models, OmpSs [4] and XiTAO [5], are precursors and testing grounds for future versions of the popular OpenMP programming model. One of the development aims of the LEGaTO project is to redesign this toolset, which was developed with performance in mind, for energy-efficient computing. The OmpSs model, coupled with its accompanying dataflow runtime (termed Nanos) and compiler (termed Mercurium), is in itself an excellent match for energy-efficient computing. OmpSs uses a pragma-based model, in which the data structures that are inputs and outputs of a particular task are identified and declared at the entrance to the task block; the inter-task dependencies are then calculated dynamically by the runtime, and a task is ready for execution when all of its dependencies have been satisfied. The OmpSs model facilitates a multitude of energy optimizations that will be implemented and refined during the project. First, since the exact data that will be used by the task is known a priori when the task is scheduled for execution; an energy-efficient data prefetching scheme will be developed to prefetch only the data that will be absolutely needed, when it is absolutely needed. Second, the tasks that are in the run queue form some kind of "lookahead" buffer enabling energy-efficient task allocation and load balancing strategies for heterogeneous hardware that will be developed during the project.

Although the task-based programming model is by itself good for energy-efficient computing on heterogeneous substrates, we

aim to further enrich the programming model and runtime for explicit support for energy-efficiency. The main idea is to attach resource requirements to parts of the computation and to execute them on dynamically constructed hardware places consisting of collections of cores and memories matching the resource annotations. Each piece of the computation is a generalized task that manages its own control flow via an embedded scheduler. Resource requirements describe the needs of the application, such as number of cores, power, reliability, and security. The tasks are annotated with the resource requirements and with their input and output structures. These annotations are propagated through the system stack for seamless integration of the software with heterogeneous hardware consisting of CPUs, FPGAs, DFEs and/or GPUs, to identify the energy-optimal execution of the task at runtime. In order to achieve this goal, the project will develop tools to determine resources based on metrics such as FLOPS/Byte, reuse distance, power consumption, etc. for individual tasks. Furthermore, the project will develop support in the programming model and runtime for heterogeneity. This will be achieved by adding topology information at the task level allowing us to select appropriate accelerators and also compute nodes in scale-out environments. A task-based programming model with a dataflow runtime is a good match for low-power hardware since tasking seamlessly enables the dispatching of processing operations close to data, while the dataflow runtime execution model is well adapted for streaming accelerators such as FPGAs or DFEs. For DFEs, the programming model currently explicitly defines where DFE execution takes place. Adding dynamic runtime support will be compelling for reasons of productivity and energy efficiency.

Finally, the LEGaTO project will apply this energy-efficient software toolset for heterogeneous hardware to three use cases. The first use case will be healthcare. The project will not only demonstrate a decrease in energy consumption in the healthcare sector; it will also show that the toolset will increase healthcare application resilience and security; both of which are critical requirements in this area. As a second use case, the project will demonstrate ease of programming and energy savings possible through the use of the LEGaTO project software–hardware framework for IoT, smart homes, and smart cities applications. Sensitive sensor information and actuator instructions can be received and sent via the developed secure IoT gateway. A third use case will be based on machine learning (ML), where the project will demonstrate how to improve energy efficiency by employing accelerators and tuning the accuracy of computations at runtime. This use case will explore object detection using Convolutional Neural Networks (CNNs) for automated driving systems and CNN- and Long Short-Term Memory (LSTM)–based methods for realistic rendering of graphics for gaming and multi-camera systems. In addition, the machine learning use case will be used to further optimize the energy efficiency in the two other use cases, as well as within the runtime.

It is important to balance the advantage of a low-energy heterogeneous CPU/FPGA/GPU hardware platform with security and resilience challenges. We are therefore working on ensuring the resilience of the software stack running on this hardware, while simultaneously optimizing for performance and low power. For fault tolerance we would like to exploit the unique characteristics of the heterogeneous CPU/GPU/FPGA platform in the runtime; for example by replicating tasks intelligently on diverse processing elements exploiting the spatial/temporal slack; additionally, we will investigate energy-efficient selective replication where only the most reliability-critical tasks will be replicated. Furthermore, we will leverage the task programming model for detecting error propagation across task boundaries and walking the task dependency graph at runtime, which will help with failure root cause analysis. Finally, we will use the properties of the task model to design application-level energy-efficient checkpointing where only the necessary and sufficient data (declared at the task entry) will be checkpointed. For security, we will develop energy-efficient security-by-design by leveraging instruction-level hardware support for security (SGX in x86 and TrustZone in ARM) to accelerate software-based security implementations.

2 About LEGaTO

The H2020 LEGaTO (Low Energy Toolset for Heterogeneous Computing) project, grant agreement n° 780681, is funded by the European Commission with a budget of more than €5 million and will last three years from its beginning on 1 December 2017. The partners of the project are Barcelona Supercomputing Center (BSC, Spain), Universität Bielefeld (UNIBI, Germany), Université de Neuchâtel (UNINE, Switzerland), Chalmers Tekniska Högskola AB (CHALMERS, Sweden), Machine Intelligence Sweden AB (MIS, Sweden), Technische Universität Dresden (TUD, Germany), Christmann Informationstechnik + Medien GmbH & Co. KG (CHR, Germany), Helmholtz-Zentrum für Infektionsforschung GmbH (HZI, Germany), TECHNION Israel Institute of Technology (TECHNION, Israel), and Maxeler Technologies Limited (MAXELER, United Kingdom).

REFERENCES

[1] ITRS. "International Technology Roadmap for Semiconductors 2.0: 2015 Edition", 2015.

[2] Ward Van Heddeghem, Sofie Lambert, Bart Lannoo, Didier Colle, Mario Pickavet, Piet Demeester, "Trends in worldwide ICT electricity consumption from 2007 to 2012", Computer Communications Volume 50, 1 September 2014.

[3] R. Griessl, M. Peykanu, J, Hagemeyer et al. "A Scalable Server Architecture for Next-Generation Heterogeneous Compute Clusters", 12th IEEE International Conference on Embedded and Ubiquitous Computing, EUC 2014

[4] A Duran, E Ayguadé, et al., "Ompss: a proposal for programming heterogeneous multi-core architectures", Parallel Processing Letters, 2011.

[5] Miquel Pericas, "ξ – TAO: A cache-centric execution model and runtime for deep parallel multicore topologies", 25th International Conference on Parallel Architectures and Compilation Techniques, PACT 2016.